Understanding the Human Genome: A Conceptual Modeling-Based Approach

(Extended Abstract)

Oscar Pastor

Centro de I+D en Métodos de Producción de Software (PROS)
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
opastor@dsic.upv.es

Nowadays, wide consensus exists about the importance of conceptual modelling in the definition of software systems that correctly fit the user's needs. In the middle of the MDD (Model-Driven Development) or MDA (Model-Driven Architecture) era, we can find a relevant set of widely spread modelling-based software production techniques, always based on well-founded concepts and methods that show that only having a precise system description, its corresponding software product can be properly understood. But it is interesting to see how these principles are not so commonly applied in current, challenging domains as Bioinformatics, where the management of tons of data is an increasing problematic issue, and where too often we can find serious problems in their manipulation that remember the data quality problems worked out for years within the Information Systems and Software Engineering communities.

This problem is often referred to as the chaos of genomic data, the main problem being that more and more are generated continuously, what makes the situation more and more complicated if we want to manage it adequately. With more than 1300 biological data sources reported to exist currently, an effective and efficient data searching strategy is a need for any activity related with the manipulation of genomic data. But the current situation makes very difficult –when not just impossible- to perform the data analysis that is required when all this huge amount of information has to be exploited. Even traditional database models that have been working successfully in the last decades are questioned when massive data analysis is needed and the volume of manipulated data is beyond the working capacity of the existing DBMS systems, normally relational-based.

In this scenario, where unstructured, low-quality data plays a prominent role, the goal of this keynote is to show how conceptual modeling techniques applied to human genome concepts can provide a working solution, by i) helping to understand and represent correctly the relevant concepts in a conceptual schema, and ii) facilitating a both rational and effective data exploitation strategy by assuring that the relevant information is properly understood from a conceptual perspective. The intention is to show that this it is the only way to enable effective management of the involved data. Only Conceptual Modeling techniques allow for providing a precise definition of relevant genomic concepts as basic as the concepts of Gene, Allele, Mutation, Splicing, Transcription Units, SNPs, Proteins... Unfortunately this view of the problem is

P. García Bringas et al. (Eds.): DEXA 2010, Part I, LNCS 6261, pp. 467–469, 2010.

not –yet!- the normal practice in current Genomic based research, where relevant concepts are managed with a high dose of conceptual uncertainty, what generates serious problems in terms of having the required data quality to manipulate it correctly.

The keynote will develop the idea that current Bioinformatics practice should be much more Information Systems (IS)-based. It is obviously true that many so-called biological ontologies, data banks and very diverse sources of genomic information exist (just as an example we could cite HUGO, Gene Ontology, Entrez Gene, Human Gene Mutation Database, VEGA, DBSnp,..), but it is also dramatically true that all these data have not a precise conceptual schema behind them, and that the data are provided in numerous, disconnected databases, that conform a set of data silos where the lack of uniformly structured data affects many basic areas, especially in the biomedical research domain. In this domain, data rely heavily on integrating and interpreting data sets produced by different experimental methods at different levels of granularity, what makes conceptual models really needed to facilitate a systematic development of biological systems.

Under this generic perspective, the benefits of a Conceptual Modeling-Oriented Genome Systems Management will be introduced, explaining how to define and exploit a Conceptual Model of the Human Genome. What problems are to be solved will be discussed, together with the main procedural steps that must conform such a CM-based process: schema definition, data loading and maintenance from different biological data sources, and data exploitation in terms of linking adequately genotype and phenotype to better understand a very old question: why we are as we are.

Instead of using a bag of information, a Conceptual Schema will introduce a kind of well-structured "cupboard", intended to assure that each piece of genomic information will be at the right place, what will make possible an efficient data exploitation, the right answers will be provided for the required questions through the achieved data integration, and the evolving nature of the genomic data will be dealt with at the right level: the conceptual one. The potential benefits of this approach will be also discussed, especially in the area of elaborating genomic reports, that is expected to open a new era in the domain of the personalized medicine.

After presenting a concrete proposal of a Conceptual Schema version for the Human Genome, the final keynote goal will be to justify its necessity, that will be based on the fact that its existence is strictly required…:

1 to enable and facilitate global research among the big set of various, distinct research groups that manipulate heterogeneous data, by fixing a conceptual gamut from which researchers can draw, in order to ensure that a 'standard dictionary of concepts' exists. This is to be achieved by:

- fixing the relevant concepts, their properties, behavior and relations, in the same way that a conceptualizing ontology would do.
- disambiguating existing biological concepts

2 to facilitate means of explicitly representing information so that:

- knowledge does not 'disappear' with it's creator
- it is clear what knowledge is present, and what is not

3 to enhance the pursued understanding of the human genome. Related to the previous comments, by looking at the selected genomic domain in terms of concepts, their

properties and relations, one disposes of a very powerful reasoning tool, useful at uncovering new relations, concepts and eventually fully understanding the given domain.

4 to drive an efficient and effective storage and processing policies for genomic data, conceptual-schema centric to assure that the well-known IS principles that enforces a data design of quality are properly considered. In that way the current manual methods applied in Bioinformatics that include tedious and repetitive tasks, with no explicit methods, prone to human errors and with a lot of required navigation through complex hyperlinks sequences, will be drastically improved.

The final objective of this work should be clear at the end of the keynote: if with Conceptual Models targeted at digital elements we have been able to improve Information Systems Development, with Conceptual Models targeted at life –as the Human Genome domain analysis makes possible- we could directly improve our living.